UNCOVERING HIDDEN INFORMATION WITHIN R&D DEPARTMENT'S TICKET USING DATA MINING CLUSTERING APPROACH

AZMI BIN ABU BAKAR

UNIVERSITI UTARA MALAYSIA

2011

UNCOVERING HIDDEN INFORMATION WITHIN R&D DEPARTMENT'S TICKET USING DATA MINING CLUSTERING APPROACH

A thesis submitted to the Faculty of Information Technology

in partial fulfillment of the requirement for the

degree Master of Science (Information Technology)

Universiti Utara Malaysia

By

AZMI BIN ABU BAKAR

KOLEJ SASTERA DAN SAINS
(College of Arts and Sciences)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certifies that)

AZMI BIN ABU BAKAR
(802332)

calon untuk Ijazah
(candidate for the degree of) **MSc. (Information Technology)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/her project of the following title)

UNCOVERING HIDDEN INFORMATION WITHIN R&D DEPARTMENT TICKET USING DATA MINING CLUSTERING APPROACH

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.
(that this project is in acceptable form and content, and that a satisfactory knowledge of the field is covered by the project).

Nama Penyelia
(Name of Supervisor) : **ASSOC. PROF. FADZILAH SIRAJ**

Tandatangan
(Signature) : _____ Tarikh (Date) : 27/2/2011

ASSOC. PROF. FADZILAH SIRAJ
Pensyarah
Bidang Sains Gunaan
Kolej Sastera & Sains
Universiti Utara Malaysia

Nama Penilai
(Name of Evaluator) : **MR. TUAN ZALIZAM TUAN MUDA**

Tandatangan
(Signature) : _____ Tarikh (Date) : 27/2/2011

TUAN ZALIZAM BIN TUAN MUDA
Lecturer
Information Technology Building
College Of Arts & Sciences
Universiti Utara Malaysia

PERMISSION TO USE

In presenting this project in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this project in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor, in her absence, by the Dean of the Faculty of Information Technology. It is understood that any copying or publication or use of this project or parts thereof for financial gain should not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my project.

Request for permission to copy or to make use of material in this project, in whole or in part should be addressed to:

Dean of the Faculty of Information Technology

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Darul Aman

ABSTRAK

Syarikat MP bahagian R&D Pulau Pinang menerima tiket setiap hari (setiap tiket yang dikemukan mewakili satu isu yang dihadapi oleh pekerjanya) dari pekerjanya yang memohon bantuan sokongan. Jumlah tahunan tiket yang dikemukan bertambah dari tahun 2007 sehingga 2009. Bertambahnya isu bermakna bertambahnya aktiviti sokongan dibuat bagi menyelesaikan isu-isu tambahan yang dihadapi. Secara langsung ini bermakna bertambahnya kos operasi. Projek ini adalah bertujuan untuk membuat analisa dengan menggunakan cara yang di deskripsi oleh teknik *Data Mining*, iaitu Clustering analysis. Maklumat tersembunyi dan sebab utama berlakunya pertambahan isu-isu tahunan dapat di bentangkan. Hasil keputusan dari analisa ini akan digunakan bagi mengolah satu kerangka ataupun penyelesaian bagi memperbaiki keadaan tersebut dan menstabilkan jumlah tiket-tiket yang dikemukakan.

Dalam kajian ini data yang diperolehi di clusterkan dengan menggunakan dua teknik data mining yang berbeza iaitu K-Mean dan Kohonen Network. Kemudian perbandingan keatas cluster-cluster yang terhasil dibuat dan dinilai dengan menggunakan Multinomial Logistic Regression dan Neural Network:MLP. Hasil keputusan menyerlahkan sebab utama (root cause) yang menyebabkan tiket-tiket dikemukakan. Maklumat ini akan digunakan oleh MP Company untuk menghasilkan kerangka ataupun kaedah penyelesaian bagi aplikasi.

ABSTRACT

MP Company's R&D department in Penang received daily submission tickets (this represent issues raised by its staffs) from it staff requesting for support. Number of annual tickets submission increases from year 2007 until 2009. Increase of issues means that increase of support activities in order to resolve these extra issues. Directly this will increase the cost of operation. This project will undergo analysis which prescribes in one of data mining technique called Clustering analysis. Hidden information and major root cause of the increase issues is expected to be unveiled. Result of this analysis can be used to generate framework or solution to improve the situation and stabilized the number of tickets submission

In this study the data extracted is clustered using two different types of data mining techniques i.e. K-Means and Kohonen Network. Later the clustered produced is compared and evaluated using Multinomial Logistic Regression and Neural Network: MLP. The result produced then reveals the biggest root caused of issue or problems that eventually triggered the ticket being submitted. This knowledge will be used by MP Company to further produce the framework or solution model for implementation.

ACKNOWLEDGEMENT

I would like to express my utmost gratitude to my supervisor Associate Professor Fadzilah Binti Siraj, whom with abundance patients, has taught and navigate me in undergoing this whole thesis.

I would also like to thanks my family who inspired, encouraged and fully supported me morally and spiritually along the way.

Last but not least, to my course mates and my friends who have helped me directly or indirectly in ensuring the completion of this research paper and tasks.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CRISP-DM	Cross-Industry Standard Process for Data Mining
DM	Data Mining
KDDM	Knowledge Discovery in Data Mining
KPI	Key Performance Index
LR	Logistic Regression
MLP	Multi-Layer Perceptron
MLR	Multinomial Logistic Regression
MP	The company whose data being analyzed
NN	Neural Network
R&D	Research and Development
SME	Small Medium Enterprise

CHAPTER 1

INTRODUCTION

In MP Company, for the Research and Development (R&D) division, a ticketing system is implemented. It is a system for user to submit a ticket from R&D staffs that use standard MP tools and applications, from hereon it referred to as users, who have request or issues their needs to be resolved internally. The issues can be ranging from *forgetting the password* up to *system development failure*.

Generically, a ticket can be viewed as a single complaint or a single request from a user to the supporting team, requesting work to be done to resolve the request or issue. Once the request is fulfilled or the issue is resolved, the responsible team will input feedback in the ticket and close the ticket. The feedback may contain work progress, details of works and status of ticket; either in progress, on hold, closed and etc.

The ticket details are stored on a centralized location in Oracle database. In this analysis the ticket extracted is filtered on tickets that have been resolved by the IT team for MP R&D in Penang. The ticket is also being taken from time to time as key

performance index (KPI) on supporting team staff, since the ticket structure can act as a log for the related staff.

In this study, the tickets were processed and analyzed using one of the data mining techniques to determine the relation of activities that contributed to the increasing number of tickets issuance from year 2007 until 2009. The technique used is called Clustering Analysis. Brief description on the process to be taken is depicted in research methods.

This study recommends a model framework for MP to implement or generate framework or system in order to reduce the increasing number of tickets issuance. A method on how to evaluate the successfulness of the implementation will also be recommended in this study.

1.1 Problem Statement

In MP Company, from Fig. 1.1, users whom are facing any kind of system failure, he or she will submit a tickets using BMC system. The support team who constantly monitor for any new ticket submitted will assist the users in resolving their issue.

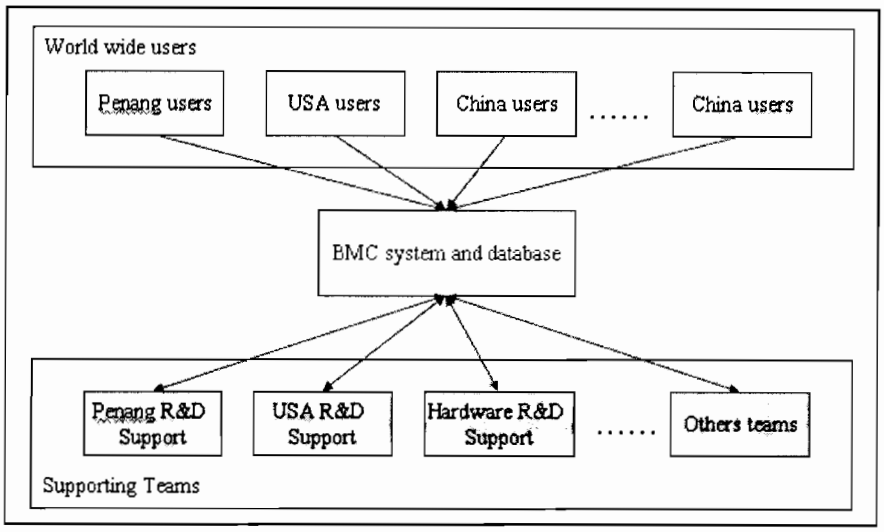


Figure 1.1: MP Company ticketing process flow

The total number of issued tickets submitted to MP R&D Penang has increase substantially. Indirectly, this leads to the rise in operation cost. Need to find the biggest root cause (issue) of the increment of tickets. This is to indirectly reduce the operation cost.

1.2 Research Objective

In general, the objective of the study is to get some explanation regarding the attribute that contribute to major cause of tickets submission

a) To find pattern that lead to one of the major root causes of the tickets in R&D MP Penang

b) To provide recommendation to MP the focus point of improvement

1.3 Research Scope

Tickets for MP R&D Penang increases within 3 years, it has increased from 2007 to 2009. Therefore, the analysis involves data from the year 2007 to 2009. An analysis that governs the major root cause of tickets submission was conducted in this study. For the purpose of clustering method, namely K-Mean and Kohonen network, once the clustered were generated, classification method such as MLR and NN were used to determine the accuracy of the clustering obtained from the clustering methods.

1.4 Significance of Study

This study is highly significant to MP R&D. Since the result enable the management to produce a plan to reduce the major root cause that occurs repeatedly. The findings can also be used by MP R&D as guide line to cluster and predict the possible incoming issue, base on the analysis of root cause.

1.5 Summary

By means of this study result will eventually enable MP Company to reduce or at least avoid increase of issues (which directly means increase of tickets submission to the support requesting for support activities).

CHAPTER 2

LITERATURE REVIEW

This chapter discussed the literature review that is related to data mining, CRISP-DM and other process flow that involved in this study.

2.1 Data Mining

Currently, data mining has been greatly acquired by business in order to improve their business activities and directly increase the profit making capability of the business. This is due that in the past recent years there's a boom in increment of data warehouse technology (Sumathi & Sivanandam, 2006)

By means of the abundance present of data, businesses implements novel technology in order to find information, particularly hidden information that enable them to excel in the highly competitive world of business. Data mining and KDDM is highly utilized by many businesses to achieve this and gain edge over the competitors. (Choinski & Chudziak, 2009)

In conjunction to this fact, data mining has attracted attention from the database communities. This is because of the data mining is applied world-wide and also it provide important technique in exploring data (Yun et al., n.d).

Another examples of data mining usage is in web page query (Cabtree et al., 2007), utilize a new interactive query expansion method, QASP. This method will determine the users query aspects and will fetch the data of websites that is clustered in according to the aspect required by users. For ambiguous query secondary input will be required from users to select the secondary finding of the query

Data mining can still be considered in it immature state. Researchers from time to time help to improvise the procedures and technique either is usage or even extending the models or standard procedures that have been establish. One of the examples is that in dealing with temporal data (Catley et al., 2009), study has indicated that extension in CRISP-DM is necessary to cater data that related to time and streaming.

2.2 Knowledge Discovery in Data Mining (KDDM)

KDDM at others time is referred as Knowledge Discovery (KD). KDDM can be simplified into steps of iterative sequence in KD. These steps of sequence are (Han & Kamber 2006): Data Cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and Knowledge presentation.

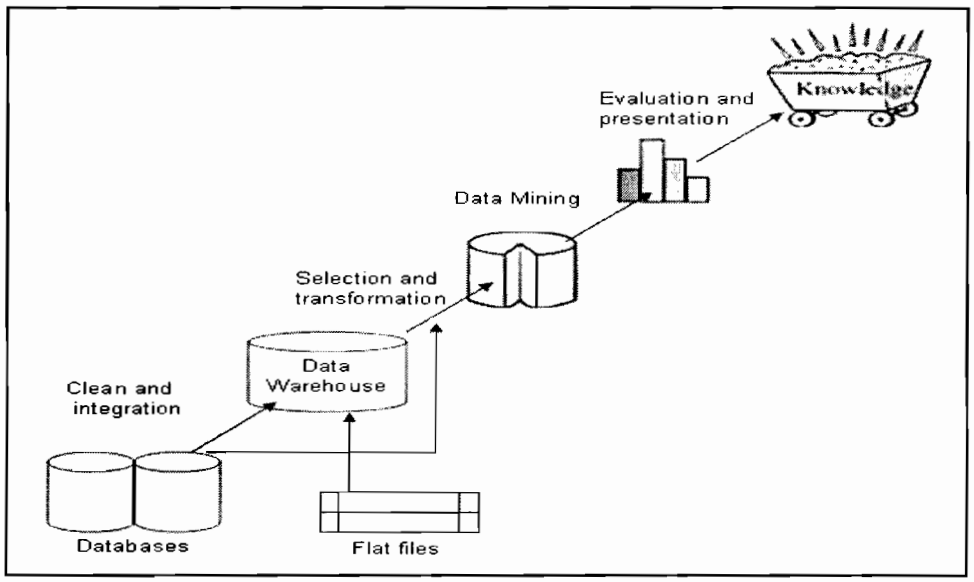


Figure 2.1 Process of Knowledge Discovery from Data Mining. (Han & Kamber, 2006)

Fig. 2.1 shows briefly the overall arrangement of the KDDM steps pictorially. The process originated from data being clean and integrated and resulted on a data warehouse format. From data warehouse the selection of data if feed to data mining for further processing. Once the result gained from the data mining process evaluation is made to ensure there is no *inconsistency* in the result. The results later on acts as knowledge that will be used in business to improve the being of the business either operationally, financially or in management.

2.3 CRISP-DM

In 1996, while interest in data mining was mounting, no widely accepted approach to data mining existed. There was a clear need for a data mining process model that

would standardize the industry and help organizations launch their own data mining projects. (Shearer, 2000)

CRISP-DM is a popular procedure that is being implemented world-wide (Zeng & Pan 2010). This model is implemented in wide range of data mining and being used in variety of purposes.

It is being used in discovering attribute of SME's behavioral pattern (Bošnjak et. al 2009). The aim is to utilize the information gained from the DM process and able to support the SME sectors in the operation and development.

In implementing CRISP-DM, the implementer is guided with generic task and output to be produced in each 6 steps of the CRISP-DM process. The details of the generic task and output are dictated in fig. 2.2 (Chapman, et. al 1999).

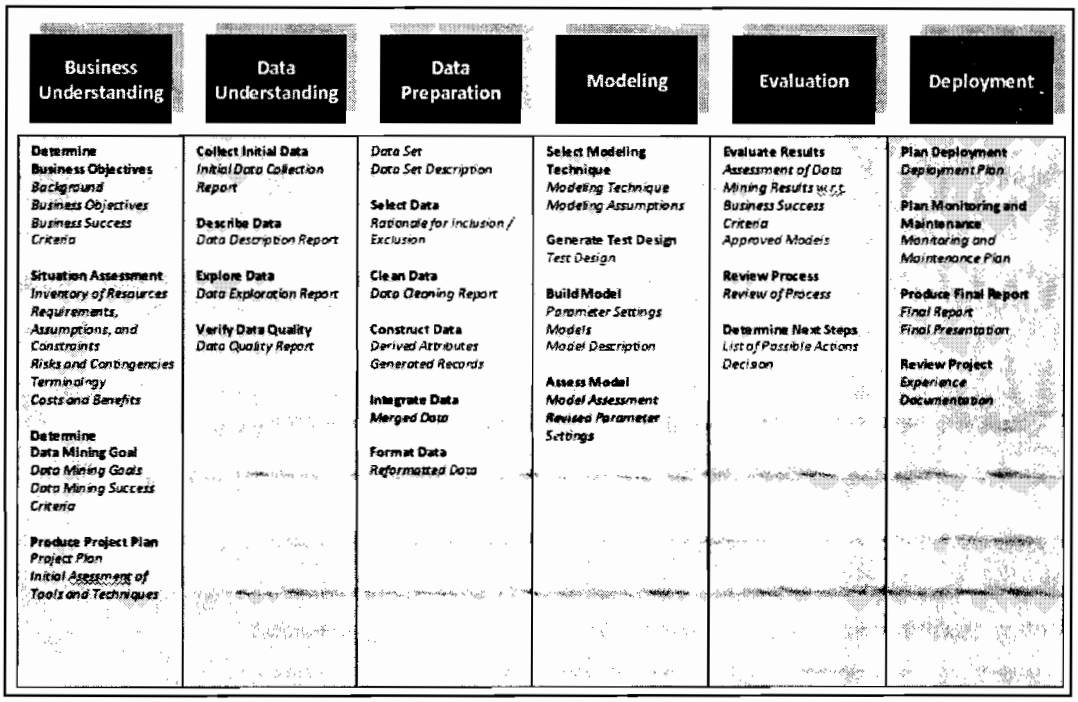


Figure 2.2: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

2.3.1 Business understanding

Every business has its objective, vision and functions. These are translated into requirement in all kind of aspects. There for in CRISP-DM, business understanding phase, is to converts all these objectives and requirement into DM problem definitions, designs and preliminary project plan (Cios et. al, 2007).

2.3.2 Data understanding

Understanding the data is crucial in data mining processing. Understand the data ensure correct processing and accuracy in the result. Data understanding process

starts from exploring the data at in the planning stage. Methods can varies in forms of interactive exploration, experimenting with the data and other subsequent processes. From here data mining task and schedule can be generated (Pan, 2009).

2.3.3 Data preparation

Stated in Fig. 2.2, data cleaning is amongst one of the process in data preparation. In order to have a reliable result and model, data cleaning has to be considered (Rahm & Do 2000). Bad quality data will affect the result of the analysis. This issue is face daily all over the world particularly with individuals or groups that involve with database and data warehousing

2.3.4 Modeling

DM descriptive approaches are identified in finding the pattern or criteria of data; the method identified is clustering analysis. Later for evaluation the predictive approaches identified, namely MLR and NN MLP (Fadzilah & Abdoulha, 2010)

a) Clustering analysis

Dealing with large data and added with the fact that there is unknown class in the particular data set, clustering analysis is the most common technique being used in data mining fields.(Yuan & Yihua 2009).

Amongst the clustering technique been implemented in this study is K-Means clustering and NN Kohonen Network.

b) Two-Step

Two-step cluster analysis has the ability to reveal the natural cluster of a data set (Webb, 2010). This feature is desirable since K-Means clustering requires the number of clustering to be input pre-processing.

c) K-Means clustering

K-Mean clustering is basically, clustering the data based mean of the members to the centroid. Setting of *centroid* is normally at random, next the calculation of the means of the data to the *centroid. Centroid* is the relocate to minimize the variance (Hastie et al., 2009)

One of the criteria of K-Means clustering is that the clustering is influenced greatly by the initial location of the *centroids*. In other word, if initial location of *centroid* if is difference, then the clustering assigned will differ on difference run of calculation. K-Means is also can be affected by any outliers data easily (Hu et al, 2009).

(Norsaadah et al., 2008), implemented Cluster analysis via K-Mean factors and Classification tree in order to produce an anthropometric sizing of Malaysia girls student of age 7 to 12 years old. The result can be used in clothing industries in making the schools uniforms for girls' student in order for the industries to be more efficient in uniform making and directly optimum the revenue.

K-Mean clustering also used to evaluate the result of clustering biological sequences in the field of drug design and disease treatment (Chen & Zhang, 2006). The clustering that generated by hierarchical agglomerative is then later evaluated via K-Mean clustering for consistency.

d) Kohonen Neural Network

Neural Network structures work with similar pattern of how human brains work, involve of neurons. However for Kohonen Neural Network differed with NN is that it consists of only input and output neurons. There is no hidden layer or bias neurons. (Heaton, 2005). It works under unsupervised mode, sometimes also known as Self Organizing Map (SOM)

2.3.5 Evaluation

Two methods in DM techniques can be used for evaluation of the result, the first method is MLR and the second method is NN MLP (Fadzilah & Abdoulha, 2010).

a) Multinomial Logistic regression (MLR)

MLR is similar to LR, they are applied when dealing with dependant variables that are nominal (Flom, n.d.). The difference is that MLR deals with calculation when the condition of the dependant variables can be more than 2 (Chan, 2005).

In this paper MLR is used to find the relationship between the dependant variables and the covariates exist (UCLA, n.d.). Namely cluster that is set and the variables that acts as the predictors

b) Neural Network

Neural network (NN) is one of the most popular data mining techniques being utilized by researchers in undergoing their analysis (SPSS Inc, 2007). NN is categorized under supervised learning in data mining.

NN is one of the predictive data mining (Ping, 2009). It is applied when application involved in predicting the results that produced by NN, and compare it with the actual data. NN is applied in various fields in prediction

NN MLP has the capability to identify the authorized user by processing the keyboard keystroke (Harun et al., 2010). From the study, NN MLP alone is able to identify the correct users of computers with 80% of accuracy. In the study NN MLP is combined with PCR and the accuracy is increased to 97%.

MLP also known as supervised network. Since it requires input and output for learn. These input and output is used in training, test and validation (NeuroDimentsion, 2010). Fig. 2.3 demonstrate the common model generated by NN MLP

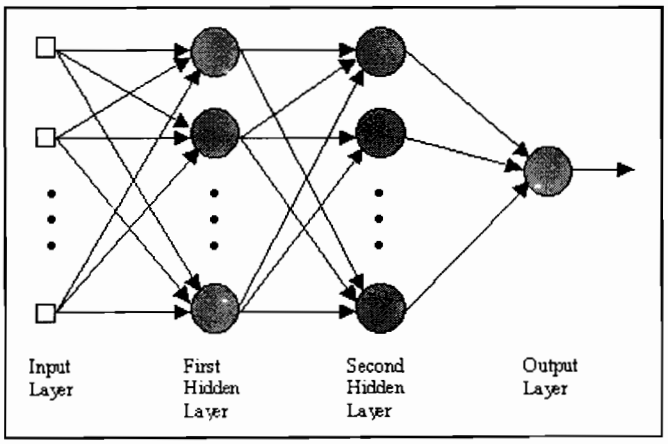


Figure 2.3: Common model of Neural Network (source NeuroDimension 2010)

2.3.6 Deployment

According to (Selbaş & Küçüksille, 2010), the deployment stages activities varies widely based on the requirement of projects and business. It can be as simple as producing a report of a project or can be as complex as implementing a solution to the problem and undergo further data mining process to monitor the results of the solution implementation.

2.4 Data Mining Hybrid Model

This model is developed with reference from CRISP-DM model (Cios et. al, 2007) the adaptation is towards academic research. Two major differences from of this model compare to the current-world-widely-accepted CRISP-DM model are that it provides wider scope of research-oriented steps. The second major amendment is that instead of modeling, in general this method introduces the data mining steps replacing it. These changes can be observed in Fig. 2.4.

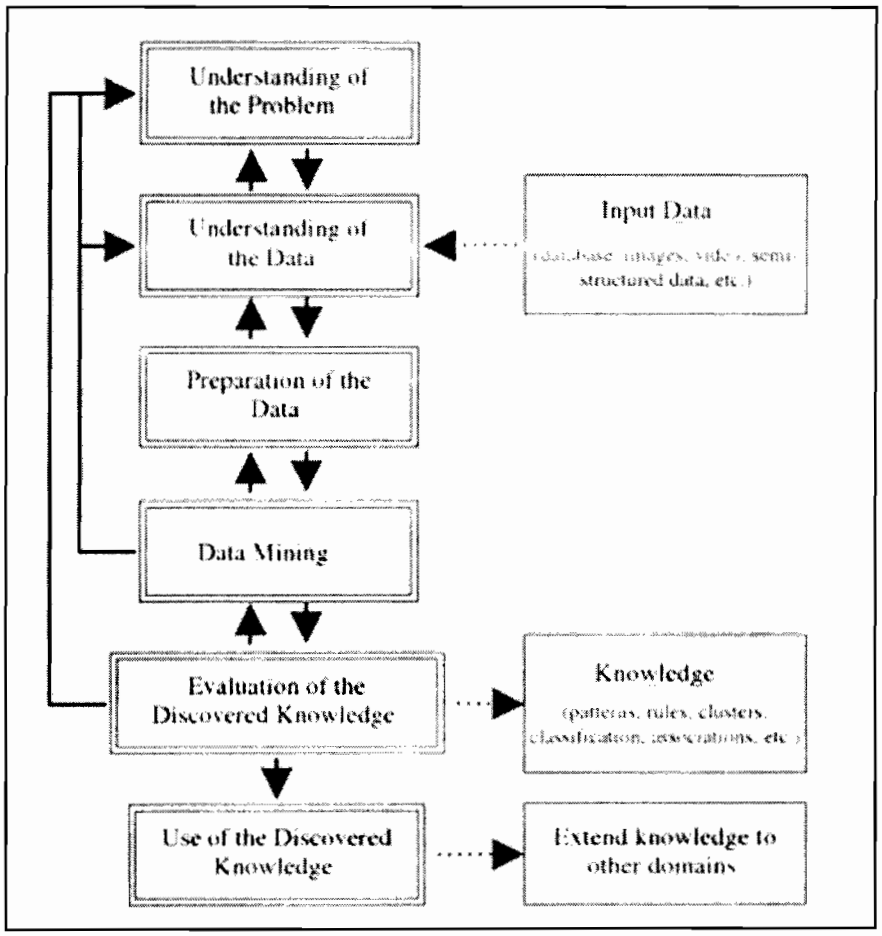


Figure 2.4: Hybrid Model source Cios et. al, 2007

CHAPTER 3

METHODOLOGY

This chapter details out DM techniques and process that was used in this study. DM is the methods that able to discover hidden data or behavior that contributed by data (Tan et al., 2006).

3.1. CRISP-DM

This data analysis and clustering process implementation will take the CRISP-DM's standard procedure as guide line. This standard procedure is widely accepted in most of the DM mining process.

Fig. 3.1 demonstrates the relation and flow of the six standard procedures surrounding the data. The steps that will act as guideline during the whole process of this thesis are business understanding, data understanding, data preparation modeling, evaluation and deployment. Taken into account as indicated that the process could be going forward or backward between the process (Chapman et al., 2000).

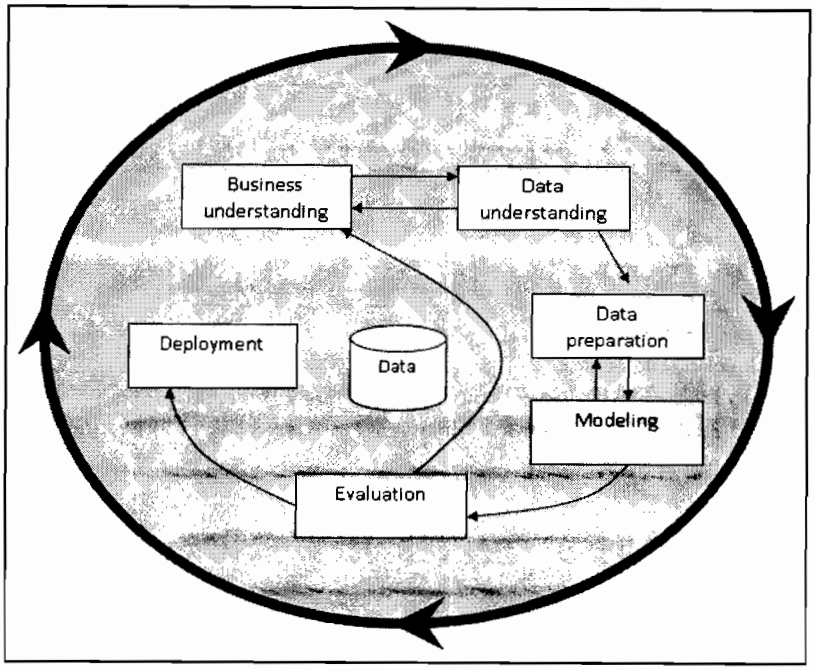


Figure 3.1: Phases of CRISP-DM Reference Model

3.1.1 Business understanding

In general, a business, regardless of it size which ranging from a multinational company to a stall at streets corner. These business activities and operation is driven by four main factors (Porter, 1979).

Fig. 3.2 reflected the four factors, which are:

1. The bargaining power of customers
2. The bargaining power of suppliers
3. The threat of new entrants
4. The threat of substitute products

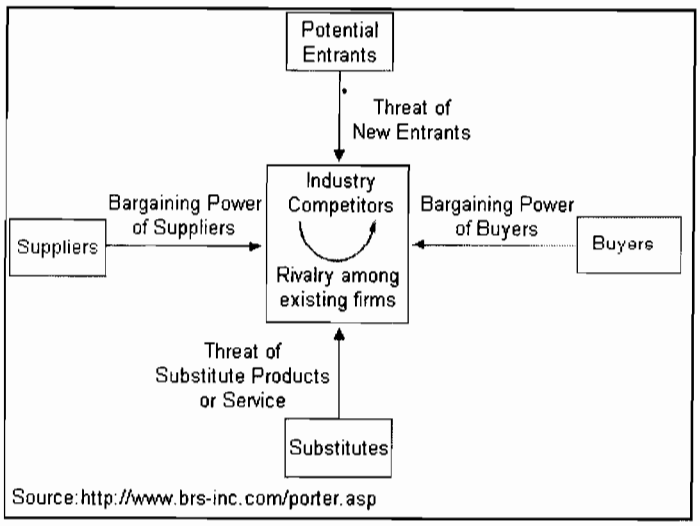


Figure 3.2: Adapted from Porter (1979)

In a telecommunication environment, it is generally known the at least more than one of these factors changes in the recent years. For example, with the increasing brands of telecommunication products, buyers bargaining power has increase. In parallel competition of product from China has also become a threat which results for business operation reviewing (Yuan & Yihua 2009). As a result a telecommunication business has to reviews back its operations activities in order to cut cost rather than to increase a product pricing which directly tend to make buyers to buy products from the competitors. Another example is that the increase of material, this also effected deeply for big business in their daily operation cost.

3.1.2 Data understanding

In a multinational telecommunication company, particularly in R&D department, when a user of a computer system facing a system, application, hardware or tools failure. The user will log a ticket requesting for supports. These tickets can be submitted via web, BMC Remedy Software (BMC Software, 2011) or even a call to their helpdesk. The tickets are then recorded into database.

The data of these tickets will represent the issues that being faces by MP R&D computers users. The data set extracted has is based on the tickets that was supported by staff from R&D Penang only.

In related to business needs and requirement, to find the major root cause of issue that the users is facing can be translated as finding the biggest cluster of tickets submitted by users and variables is the category of operations and products that included in tickets.

By exploration, it is decided that the 6 field is significant to be used in this analysis process.

The fields used are:

a) OPERATIONAL_TIER_ONE

b) OPERATIONAL_TIER_THREE

c) OPERATIONAL_TIER_TWO

d) PRODUCT_TIER_ONE

e) PRODUCT_TIER_THREE

f) PRODUCT_TIER_TWO.

3.1.2 Data preparation

The data extracted was analyzed for its locality of the users that requested for support from Penang support team.

Total of ticket number is 662. From Table 3.1 the highest frequency of tickets is from Malaysia it self.

Table 3.1: Frequency of ticket by location of submitters

	Frequency	%
United States	4	0.60%
Argentina	1	0.15%
Malaysia	634	95.77%
China	19	2.87%
India	3	0.45%
Italy	1	0.15%
Total	662	

From the bar chart in Fig. 3.3, clearly can be concluded that in the dataset, 95.77% of the tickets is submitted by users from Malaysia, while the others is from United States of America, Argentina, China, India and Italy.

Therefore for further analysis onward, the tickets from Malaysia only will be analyzed since it shows immense significance difference. The ticket submitted by Malaysia's users alone total up to 634 tickets.

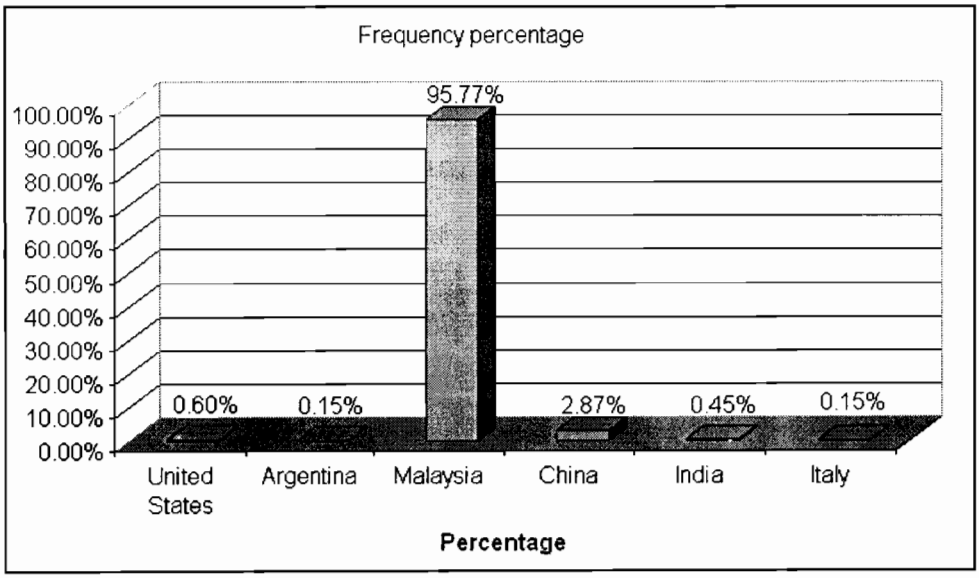


Figure 3.3: Graph of percentage of ticket submission by location

Next the variables that are used are in description manner, which is as string. Therefore 6 new variables are created in order to replace the 6 original variables.

The variables created are OPERATION1, OPERATION2, OPERATION3, PRODUCTION1, PRODUCTION2 and PRODUCTION3. The value mapping between the original sets of variable and the value in new variable can be referred in Appendix A.

3.1.3 Modeling

In process of finding the major root cause, a few modeling technique of data mining utilized to achieve the valid major cluster of tickets.

To implement the modeling of data mining, SPSS 16.0 and Clementine 12.0 is used to process the data. SPSS 16 usage is on Two-Step analysis, K-Means clustering, MLR and Neural Network MLP. Clementine 12.0 is used to process the dataset Kohonen Clustering.

Clustering techniques is very useful in finding the categorical type of data. (Ghaderi et al., 2007), applies clustering technique to derive the pattern of electricity utilization. The study uses Maple software to factor out the variable of environmental parameters. The result of the study can be use in Macro Economics Decision Making and Middle Management Technical Engineering, in planning the tasks.

Listed below are the details of task undergone in each of the modeling techniques used:

a) Two-Step

The sole purpose of this step is to find the best number of cluster to be used. Two-step provides the best number of clusters by balancing between BIC changes and the Ratios of Distance Measures.

Processing of the data set using two-step as follows:

i. Data is loaded on the SPSS

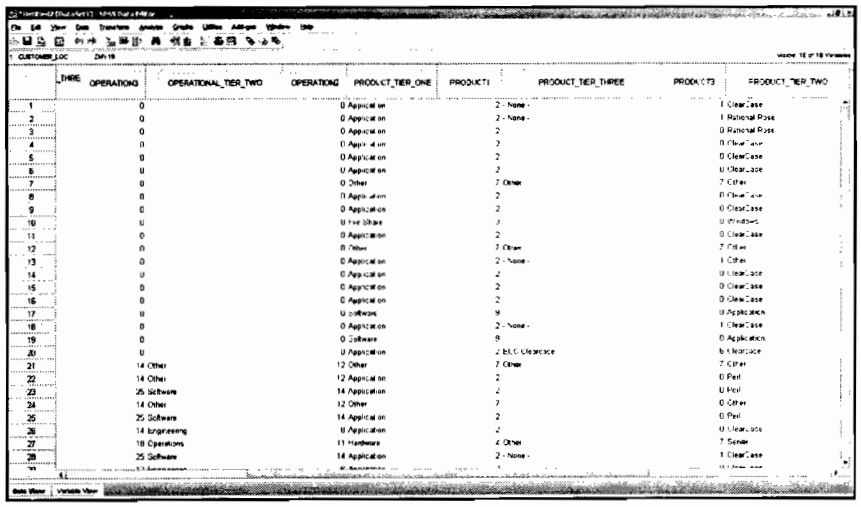


Figure 3.4 Data set loaded into SPSS application

ii. Select menu option as shown if Fig. 3.5

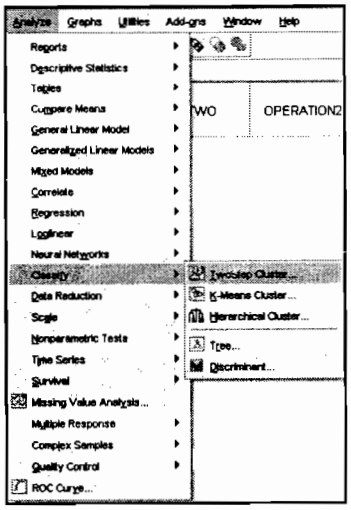


Figure 3.5 Menu option chosen for two-step

iii. The variables selected for two-step process as shown is Fig. 3.6.

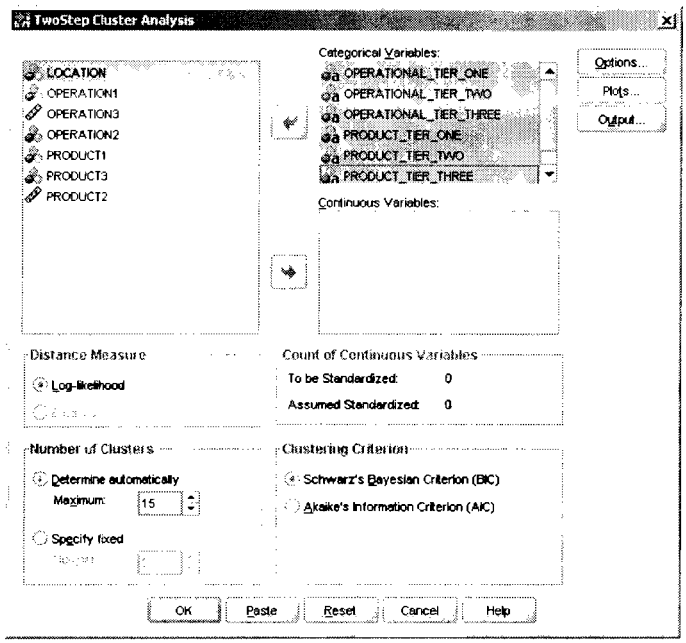


Figure 3.6 Variable selected for Two-Step processing

iv. The option selected in the Plot option is shown in Fig. 3.7

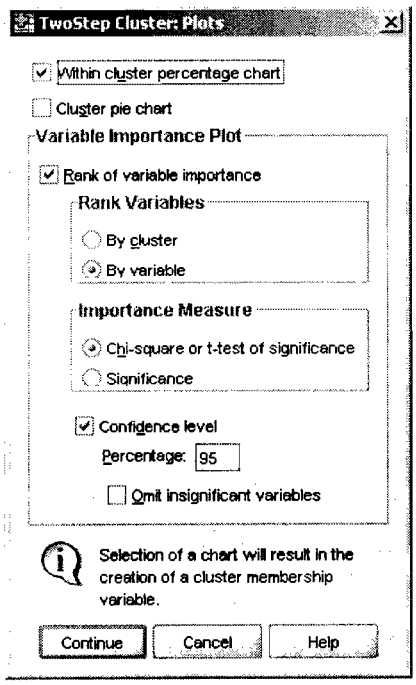


Figure 3.7 The option selected in the *Plot* option

From Table 3.2 the Two-Step processing auto select 4 as the best cluster number to be implemented for this dataset.

Table 3.2: Result of auto-clustering via Two-Step

Auto-Clustering

Number of...	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	12765.401			
2	11410.823	-1354.578	1.000	1.305
3	10554.396	-856.427	.632	1.568
4	10288.821	-265.575	.196	2.361
5	10622.625	333.804	-.246	1.067
6	10983.941	361.316	-.267	1.222
7	11420.168	436.228	-.322	1.085
8	11882.942	462.774	-.342	1.170
9	12390.943	508.001	-.375	1.128
10	12929.247	538.304	-.397	1.028
11	13474.012	544.765	-.402	1.080
12	14035.772	561.760	-.415	1.095
13	14616.043	580.271	-.428	1.061
14	15207.514	591.471	-.437	1.194
15	15828.704	621.190	-.459	1.169

b) K-Means clustering

From Two-Step analysis, 4 is the best cluster number for implementation. K-Means clustering is then triggered to process the data set to achieve the clustering membership

Step to process the data set via K-Means clustering as follow:

i. Load the data into SPSS application

ii. Select the menu as shown in Fig. 3.8

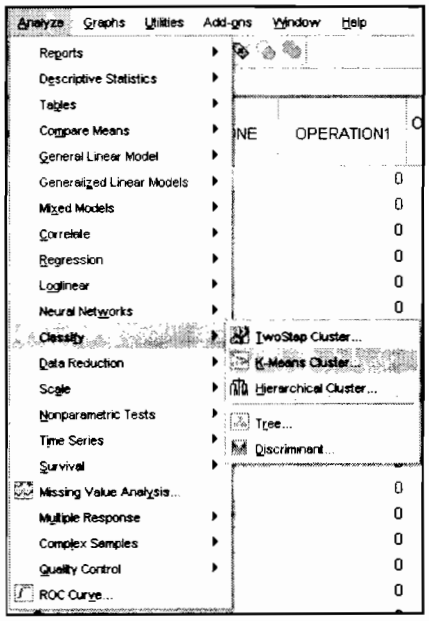


Figure 3.8 Menu selections on K-Means clustering

iii. Increase the number of iteration to 20 as show in Fig. 3.9. This is to ensure the K-Means centroids are stable once the iteration loops is complete.

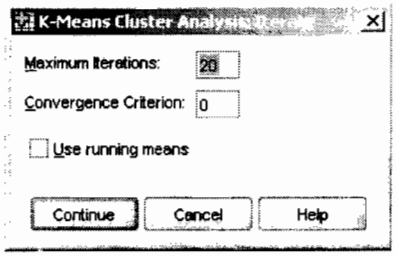


Figure 3.9 Increase the number of iteration

iv. The option selected on option prior the K-Means clustering calculation processes is shown in Fig. 3.10.

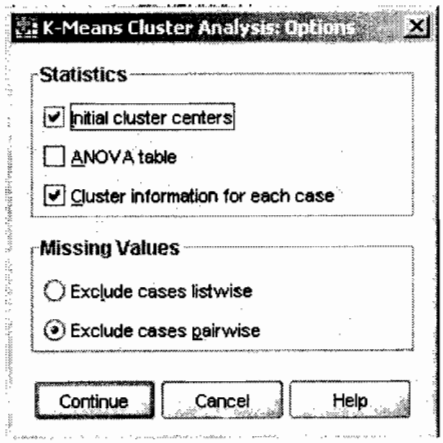


Figure 3.10 The option selected K-means clustering.

Table 3.3: Show the assignment of cluster to data using K-Means clustering (cut-off-at center)

Cluster Membership

Ca...	Cluster	Distance
1	1	11.026
2	4	15.455
3	4	16.517
4	1	11.005
5	1	11.005
6	1	11.005
7	4	17.130
8	1	11.005
9	1	11.005
10	4	19.621
11	1	11.005
12	4	17.130
13	4	15.939
⋮	⋮	⋮
627	4	15.918
628	4	14.042
629	2	2.974
630	4	15.783
631	2	2.845
632	4	15.161
633	2	8.169
634	2	4.847

By compiling the data from Table 3.3, with the assignment of cluster to each data row, a figure can be generated to obtain criteria of each cluster together with the frequency of cluster.

c) Neural Network: Kohonen Network

In order to generate cluster membership from the data set, Clementine is utilized to process the data. The cluster assigned to data row once again used to generate the criteria of cluster and frequency.

3.1.4 Evaluation

Once the cluster is set to the data set, evaluation of the cluster is necessary to ensure that the clustering is valid on different Modeling techniques. To achieve this two technique of evaluation is selected; i.e. MLR and NN MLP

Evaluation can also be made by observing the two cluster criteria generated via K-Mean and Kohonen clustering.

a) Multinomial Logistic Regression (MLR)

The clustering assignment that has been set then is feed to MLR processing to generate the relationship and model of the attributes involved with the dependant variables.

b) Neural Network: MLP

Analyzing using NN MLP, we can determine the accuracy (Zhou, 2004) of the clustering that has been set by Kohonen clustering. We can evaluate the clustering that was set from the Kohonen clustering.

3.1.5 Deployment

The result gained from this study will be presented to MP Company. With the knowledge, model of solution or process flow framework can be generated in order to proactively avoid the issues from happening.

CHAPTER 4

RESULT

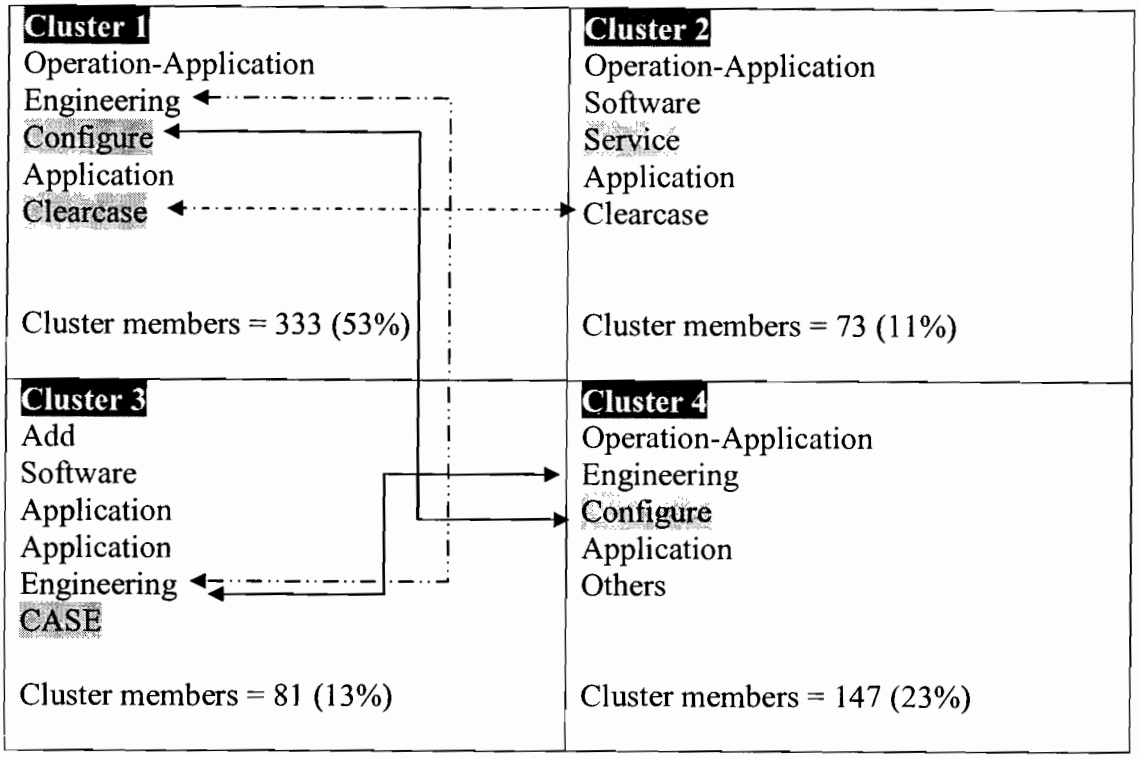
The results of the study are reported in this chapter. The discussions are on the result gained from the entire test mentioned in previous section.

4.1 K-Means Clustering

K-Means clustering is a centroids based clustering (Matteucci, n.d.). The data set is calculated to measures the closest centroid in loop. The centroids location is changed as necessary with relation to its members. The loops stop when the centroids stop to change its location.

The results from K-Means clustering are shown in Table 4.1. A total of 634 cases involved in this study. The tickets attribute with respect to clusters are listed in Table 4.1.

Table 4.1: Criteria of cluster (K-Means)



A few rules could be extracted from the 4 clusters such as:

IF ticket = *Service* then the tickets is in Cluster 2

IF ticket = *Engineering* then the ticket is in Cluster 1, 3, 4 but not 2

IF tickets = *Configure* then the tickets is in Cluster 1 or 4

IF tickets = *Clearcase* then the ticket is in Cluster 1 or 2

One interesting observation to note is that for all clusters, *Application* appears in them.

From Fig. 4.1, out of 634, the highest percentage of cluster membership is Cluster 1 (53%), Cluster 4 (23%), Cluster 3 (13%) and Cluster 2 (11%).

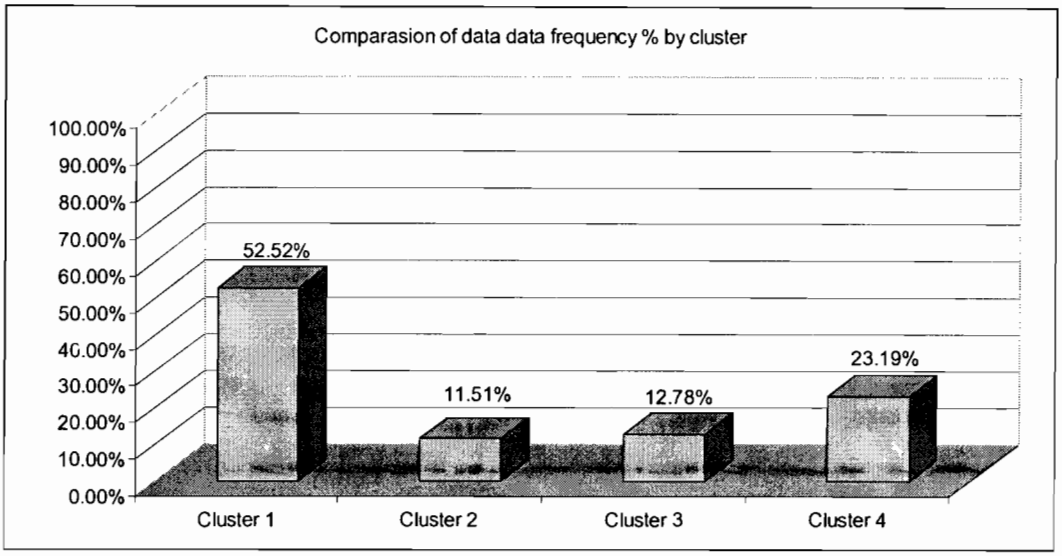


Figure 4.1: Cluster Frequency percentage in graphical mode

To distinguish between clusters the rules can be written as:

IF ticket = *Clearcase* AND ticket = *Configure* then Cluster = 1

IF ticket = *Configure* AND ticket = *Others* then Cluster = 4

IF ticket = *Add* OR ticket = *CASE* then Cluster = 3

Table 4.2: Spearman Correlation (K-Means)

Correlations

			LOCATION	OPERATION1	OPERATION2	OPERATION3	PRODUCT1	PRODUCT3	PRODUCT2	CLUSTER
Spearman's rho	LOCATION	Correlation Coefficient	.	.	.	.	.	.	.	.
		Sig. (2-tailed)	.	.	.	.	.	.	.	.
		N	634	634	634	634	634	634	634	634
	OPERATION1	Correlation Coefficient	.	1.000	.178**	.195**	.257**	-.027	.129**	.162**
		Sig. (2-tailed)	.	.	.000	.000	.000	.496	.001	.000
		N	634	634	634	634	634	634	634	634
	OPERATION2	Correlation Coefficient	.	.178**	1.000	-.026	.138**	.441	.206	.283**
		Sig. (2-tailed)	.	.000	.	.511	.000	.000	.000	.000
		N	634	634	634	634	634	634	634	634
	OPERATION3	Correlation Coefficient	.	.195	-.026	1.000	.159**	-.031	-.067	.142**
		Sig. (2-tailed)	.	.000	.511	.	.000	.441	.091	.000
		N	634	634	634	634	634	634	634	634
	PRODUCT1	Correlation Coefficient	.	.257**	.138**	.159**	1.000	.156**	.095*	.266**
		Sig. (2-tailed)	.	.000	.000	.000	.	.000	.017	.000
		N	634	634	634	634	634	634	634	634
	PRODUCT3	Correlation Coefficient	.	-.027	.441**	-.031	.156**	1.000	.153**	.296**
		Sig. (2-tailed)	.	.496	.000	.441	.000	.	.000	.000
		N	634	634	634	634	634	634	634	634
	PRODUCT2	Correlation Coefficient	.	.129**	.206**	-.067	.095*	.153**	1.000	.650**
		Sig. (2-tailed)	.	.001	.000	.091	.017	.000	.	.000
		N	634	634	634	634	634	634	634	634
	CLUSTER	Correlation Coefficient	.	.162**	.283**	.142**	.266**	.296**	.650**	1.000
		Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.000	.
		N	634	634	634	634	634	634	634	634

From Table 4.2, we can clearly says that there's highly significant relation between Cluster and Product2 ($p = 0$, $r = 0.85$) and medium relation for Cluster and OPERATION2, PRODUCT1 and PRODUCT2 with ($p = 0$, $r = 0.283$), ($p = 0$, $r = 0.266$) and ($p = 0$, $r = 0.296$) respectively.

4.2 Neural Network: Kohonen Network

Kohonen Network one of the neural network that trained under unsupervised learning (Wikipedia, 2011). It produces a two dimensional map, with reference of this study clusters.

Table 4.3 displays the criterion of the Clusters generated when the data set is process via Kohonen Network clustering

Table 4.3: Criteria of cluster (Kohonen)

Cluster 1	**Cluster 2**
Operation-Application Engineering Configure Application Clearcase Cluster members = 297 (47%)	Operation-Application Business Failure Application Clearcase Cluster members = 82 (13%)
Cluster 3	**Cluster 4**
Add Software Application Application Engineering CASE Cluster members = 136 (21%)	Operation-Application Software Service Application Clearcase Cluster members = 119 (19%)

The first step is to compare with the Cluster criteria generated via K-Mean clustering.

By comparison it can be observer that

1. Cluster 1 (generated via Kohonen) has the same criteria with Cluster 1 from K-Mean clustering.
2. Both clustering methods furnish their Cluster 1 with the highest number of membership, even though not with the same frequency.
3. Cluster 3 (generated via Kohonen) has the same criteria with Cluster 3 from K-Mean clustering.
4. Cluster 4 from Kohonen Clustering has the same criteria with Cluster 2 from K-Mean clusters

From Fig. 4.2, out of 634, the highest percentage of cluster membership is Cluster 1 (47%), Cluster 3 (21%), Cluster 4 (19%) and Cluster 2 (13%).

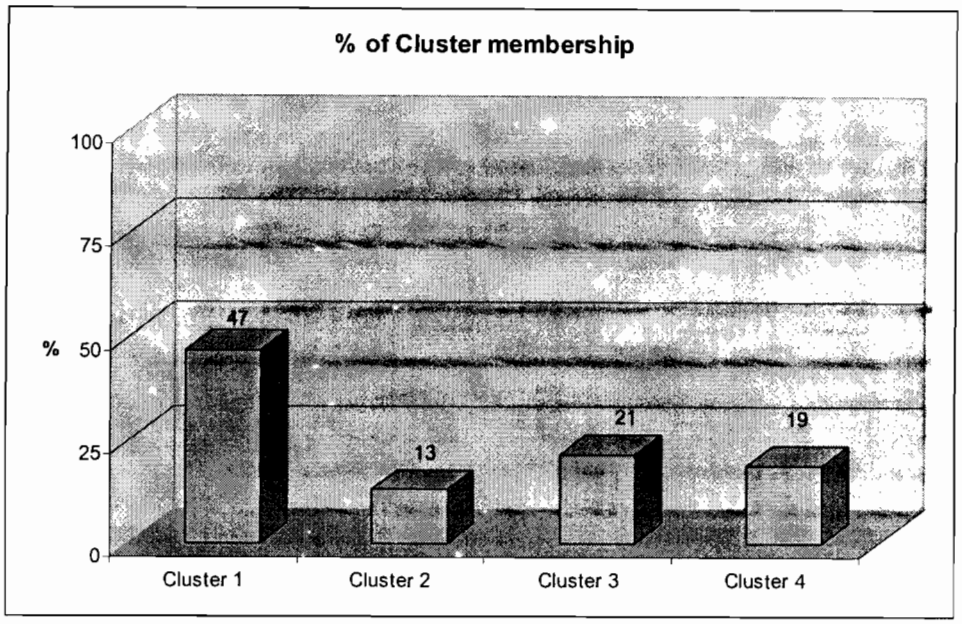


Figure 4.2: Percentage of Cluster members via Kohonen Network

In Table 4.4, is the result of correlation coefficient to determine the relation between attribute OPERATION1, OPERATION2, OPERATION3, PRODUCT1, PRODUCT2 and PRODUCT3 with the cluster.

Table 4.4: Spearman Correlation (Kohonen)

	OPERATION1	OPERATION2	OPERATION3	PRODUCT1	PRODUCT2	PRODUCT3
Correlation Coefficient	0.158	0.310	0.077	0.193	0.113	0.587
Sig. (2-tailed)	0.000	0.000	0.051	0.000	0.004	0.000

The highest correlation is between Cluster and PRODUCT3 (p=0.00, r = 0.587), the next highest is between Cluster and OPERATION2 (p=0.00, r = 0.310)

4.3 Multinomial Logistic Regression (MLR)

MLR is used in this evaluation is due to that the result of clustering has 4 levels, and the dependant variable is nominal. The application used in this process is SPSS 16.0 and on Microsoft Windows XP SP3.

Pre processing the data using MLR, frequency test is done to get the highest cluster membership. From the frequency test result shown in Table 4.5, cluster 1 has the most number of tickets; therefore in processing the data via MLR, cluster 1 will be taken as reference cluster.

Table 4.5: Result from frequency test.

Cluster					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	297	46.8	46.8	46.8
	2	82	12.9	12.9	59.8
	3	136	21.5	21.5	81.2
	4	119	18.8	18.8	100.0
	Total	634	100.0	100.0	

In modeling MLR two steps will be taken. In the first all variables i.e OPERATION1, OPERATION1, OPERATION1, PRODUCT1, PRODUCT2, and PRODUCT3 will be included in modeling using MLR.

In the second step only OPERATION2 and PRODUCT3 will be taken. This is to evaluate the significant that can be observed in Table 4.2 with value 0.443.

4.3.1 Using all variable as covariate

Table 4.6: Result of beta value from MLR

Cluster		B	Wald	Sig.
2	Intercept	-1.34	5.49	0.019
	OPERATION1	0.97	17.84	0.000
	OPERATION3	0.14	25.92	0.000
	OPERATION2	-0.63	99.38	0.000
	PRODUCT1	0.05	0.21	0.644
	PRODUCT3	20.03	36715.33	0.000
	PRODUCT2	0.02	0.51	0.475
3	Intercept	-6.47	65.42	0.000
	OPERATION1	1.18	34.45	0.000
	OPERATION3	-0.16	15.15	0.000
	OPERATION2	-0.15	4.07	0.044
	PRODUCT1	0.70	51.84	0.000
	PRODUCT3	20.36	49313.94	0.000
	PRODUCT2	0.10	16.03	0.000
4	Intercept	-2.86	27.33	0.000
	OPERATION1	1.02	26.15	0.000
	OPERATION3	0.10	16.50	0.000
	OPERATION2	-0.33	28.87	0.000
	PRODUCT1	0.13	1.83	0.176
	PRODUCT3	19.99	.	.
	PRODUCT2	0.01	0.23	0.629

a. The reference category is: 1.

In relation to Cluster 1, when the alpha is set to 0.05. The model generated for each cluster reject null hypothesis for these variables, which has significant value < 0.05. List of variable for each cluster models as follows:

For Cluster 2

OPERATION1

OPERATION3

OPERATION2

PRODUCT3

For Cluster 3

OPERATION1

OPERATION3

OPERATION2

PRODUCT1

PRODUCT3

PRODUCT2

For Cluster 4

OPERATION1

OPERATION3

OPERATION2

Base on Table 4.6 model generated:

Cluster 2

[1/1-p] = exp (-1.34 + (0.97)*OPERATION1 + (-0.63)*OPERATION2 + (0.14)*OPERATION3 + (0.02)*PRODUCT3)

Cluster 3

[1/1-p] = exp (-6.47 + (1.18)*OPERATION1 + (-0.15)*OPERATION2 + (-0.16)*OPERATION3 + (0.7)*PRODUCT1 + (0.1)*PRODUCT2 + (20.36)*PRODUCT3)

Cluster 4

[1/1-p] = exp ((-2.86)*OPERATION1 + (-0.33)*OPERATION2 + (0.1)* OPERATION3)

Table 4.7: Classification table and prediction

Classification

Observed	Predicted				
	1	2	3	4	Percent Correct
1	292	0	4	1	98.3%
2	4	49	6	23	59.8%
3	8	6	117	5	86.0%
4	40	23	7	49	41.2%
Overall Percentage	54.3%	12.3%	21.1%	12.3%	80.0%

From Table 4.7 the prediction highest correct percentage is on Cluster 1 (98.3%), followed by Cluster 3 (86%), Cluster 2 (59.8%) and Cluster 4 (41.2%)

4.3.2 Using OPERATION2 + PRODUCT3 as covariate

In relation to Cluster 1, when the alpha is set to 0.05. The model generated for each cluster reject null hypothesis for these variables, which has significant value < 0.05. List of variable for each cluster models as follows:

a) For Cluster 2

OPERATION2

PRODUCT3

b) For Cluster 3

OPERATION2

PRODUCT3

c) For Cluster 4

<No variables rejected>

Table 4.8: Result of beta value from MLR

Cluster		B	Wald	Sig.
2	Intercept	0.30	0.93	0.334
	OPERATION2	-0.31	47.95	0.000
	PRODUCT3	17.36	27370.35	0.000
3	Intercept	-4.00	63.41	0.000
	OPERATION2	0.18	13.70	0.000
	PRODUCT3	17.69	77423.88	0.000
4	Intercept	-1.89	27.81	0.000
	OPERATION2	0.05	1.48	0.224
	PRODUCT3	17.35	.	.
a. The reference category is: 1.				

*** All variable is significant since all p<0.05**

d) Model generated

Cluster 2 = [1/1-p] = exp (0.3 + (-0.31)*OPERATION2 + (17.36)*PRODUCT3)

Cluster 3 = [1/1-p] = exp (-4 + (0.18)*OPERATION2 + (17.69)*PRODUCT3)

Cluster 4 = [1/1-p] = exp (-1.89)

Table 4.9: Classification table and prediction

Classification

Observed	Predicted				
	1	2	3	4	Percent Correct
1	297	0	0	0	100.0%
2	53	0	4	25	.0%
3	41	9	85	1	62.5%
4	39	32	47	1	.8%
Overall Percentage	67.8%	6.5%	21.5%	4.3%	60.4%

From Table 4.9 the prediction highest correct percentage is on Cluster 1 (100.0%), followed by Cluster 3 (62.5%), Cluster 4 (0.8%) and Cluster 2 (0%)

4.4 Neural Network: MLP

The application used in this process is SPSS 16.0 and on Microsoft Windows XP SP3.

During this process, 3 sets of parameters setting has been use in this MLP analysis. The parameters sets details are:

4.4.1 Parameters set 1 (This is SPSS default setting)

Training	70%
Test	30%
Hold out	0%

Result as follows:

Table 4.10: Classification table and prediction (Parameters set 1)

Classification

Sample	Observed	Predicted 1	2	3	4	Percent Correct
Training	1	198	0	0	0	100.0%
	2	0	54	0	0	100.0%
	3	0	0	93	0	100.0%
	4	0	0	0	83	100.0%
	Overall Percent	46.3%	12.6%	21.7%	19.4%	100.0%
Testing	1	96	0	0	0	100.0%
	2	1	25	0	0	96.2%
	3	0	0	39	0	100.0%
	4	0	1	0	32	97.0%
	Overall Percent	50.0%	13.4%	20.1%	16.5%	99.0%

Dependent Variable: Cluster

Observed from Table 4.10 using default parameters in predicting the data set with Training = 70% and Test = 30%. On the training phase NN manage get 100% accuracy on predicting all clusters. Later in the Testing phase the prediction drops for Cluster 2 and Cluster 4 with 96% and 97% respectively.

Fig. 4.3 is the model of NN MLP generated. It dictated 3 hidden layer neurons and a bias neuron. At the input layer there is also a bias neuron.

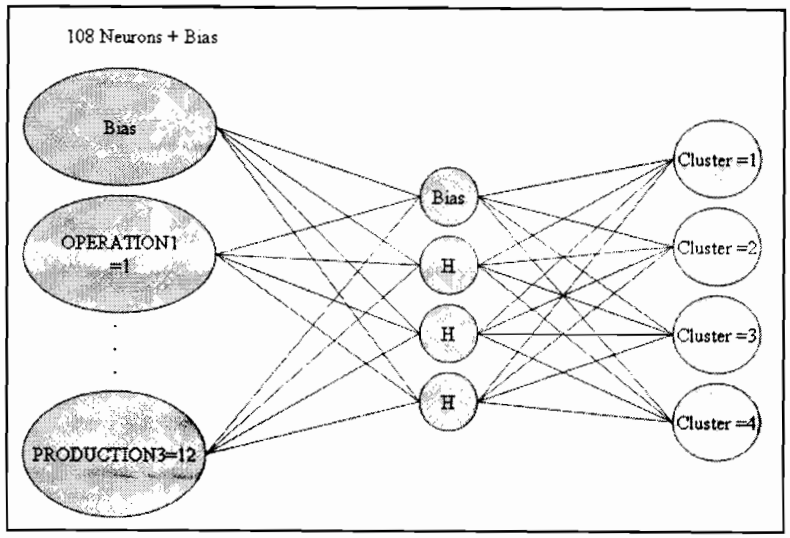


Figure 4.3: NN network generated with 3 neuron of hidden layer and 2 bias neurons.

4.4.2 Parameters set 2

Training	80%
Test	10%
Hold out	10%

Result as follows:

Observed from Table 4.11 using default parameters in predicting the data set with Training = 80% and Test = 10% and validation = 10%. On the training phase NN manage get 100% accuracy on predicting all clusters. In the testing phase NN also managed to get 100% for all cluster. This is NN have larger number of data in training. Later in the Holdout phase the prediction drops for Cluster 4 with 93%

Table 4.11: Classification table and prediction (Parameters set 2)

Classification

Sample	Observed	Predicted 1	2	3	4	Percent Correct
Training	1	240	0	0	0	100.0%
	2	0	64	0	0	100.0%
	3	0	0	115	0	100.0%
	4	0	0	0	94	100.0%
	Overall Percent	46.8%	12.5%	22.4%	18.3%	100.0%
Testing	1	26	0	0	0	100.0%
	2	0	7	0	0	100.0%
	3	0	0	12	0	100.0%
	4	0	0	0	10	100.0%
	Overall Percent	47.3%	12.7%	21.8%	18.2%	100.0%
Holdout	1	29	0	0	0	100.0%
	2	0	10	0	0	100.0%
	3	0	0	8	0	100.0%
	4	0	0	1	14	93.3%
	Overall Percent	46.8%	16.1%	14.5%	22.6%	98.4%

Fig. 4.4 is the model of NN MLP generated. It dictated 4 hidden layer neurons and a bias neuron. At the input layer there is also a bias neuron.

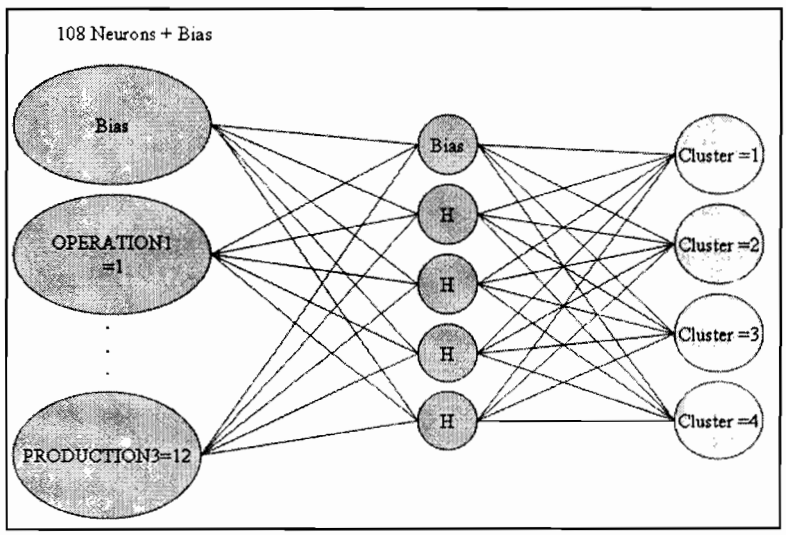


Figure 4.4: NN network generated with 4 neuron of hidden layer and 2 bias neurons.

CHAPTER 5

CONCLUSION

This chapter provides the summary and conclusion based on the findings from the data mining test applied to the data set.

5.1. Conclusion

The test results clearly demonstrate to us, that Cluster 1 and cluster 3 is consistent with the test of K-Means clustering and Kohonen Network. Added with the predictive test using MLR and NN MLP, we achieved that Cluster 1 and Cluster 3 the highest of accuracy amongst the other cluster.

We can also observe from the Criteria of the cluster extracted from the two clustering methods that is K-Means and Kohonen Network, from the two criteria, cluster 1 and cluster 3 has the exact same criteria on both methods. While for Cluster 2 and Cluster 4 it is either substituted or interchangeable.

If we compare between Cluster 1 and cluster 3, from the frequency we can see that from the K-Means clustering, cluster 1 has higher number of membership of 53%

compare to Cluster 3 with only 13% and Cluster 3 percentage is lower than Cluster 4.

Where as for the Kohonen Network Clustering, Cluster 1 similar to K-Means has higher membership of 47%, and Cluster 3 has 21% but this time it is higher than Cluster 4.

All in all, with regard to the objective of the study is to find the major root cause of the issue being face by MP Company in the R&D department. It can be said that from the ticket the major root cause is regarding *Operation* on *application* on *engineering* activities which involving *Clearcase application configuration*.

5.2 Recommendation

Clearcase is a software that provides complex version control of application development and source coding. It also caters in matters of workspace management, parallel development and builds auditing for productivity improvement (IBM Corp. 2010).

It is recommended that framework or solution module be generated surrounding the issue of Clearcase configuration on the engineering application usage.

REFERENCE

BMC Software (2011). Retrieved 21 Feb 2011 from http://www.bmc.com/products

Bošnjak, Z., Grljević, O, & Bošnjak, S. (2009). *CRISP-DM as a Framework for Discovering Knowledge in Small and Medium Sized Enterprises' Data. 5th International Symposium on Applied Computational Intelligence and Informatics held on 28 May–29 May 2009 at* Timişoara, Romania

Catley, C., Smith, K., McGregor, C., & Tracy, M. (2009). *Extending CRISP-DM to Incorporate Temporal Data Mining of Multidimensional Medical Data Streams: A Neonatal Intensive Care Unit Case Study.* University of Ontario Institute of Technology, Oshawa, Canada

Chan, Y.H. (2005). *Basic Statistics For Doctors: Biostatistics 305. Multinomial logistic regression.*(CME Article). Singapore Med J.

Choinski, M., & Chudziak, J. A. (2009) Proceedings of the International Multiconference on Computer Science and Information Technology: *Ontological Learning Assistant for Knowledge Discovery and Data Mining.* pp. 147–155

Cios, K. J., Pedrycz, W., & Swiniarski, W. (2007). *Data Mining: A Knowledge Discovery Approach,* New York: Springer Science + Business Media, LLC

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Thomas, R., Shearer, C., & Wirth, R., (2000). *CRISP-DM 1.0 Step-by-step data mining guide.*

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Thomas, R., Shearer, C., & Wirth, R., (1999). *The CRISP-DM Process Model.* CRISP-DM Discussion Paper.

Chen, W.B., and Zhang C., (2006). *A Robust Method for Biological Sequence Clustering.* Alabama: University of Alabama

Crabtree, D., Andreae, P. & Gao, X., (2007). *Understanding Query Aspects with applications to Interactive Query Expansion*, New Zealand: Victoria University of Wellington

Gashi, I., Stankovic V., Leita, C. and Thonnard, O., (2009). *An Experimental Study of Diversity with Off-The-Shelf AntiVirus Engines,* London: City University London

Ghaderi, S.F., Azadeh, A., & Keyno, H. S. (2007). *Forecasting electricity consumption by separating the periodic variable and decompositions the pattern,* Iran: University of Tehran

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Technique* (2nd ed.). San Francisco: Morgan Kaufmann Publisher

Harun, N., Dlay, S.S., & Woo, W.L. (2010). *Performance of Keystroke Biometrics Authentication System Using Multilayer Perceptron Neural Network (MLP NN).* United Kingdom: Newcastle University.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer Science + Business Media, LLC

Heaton, J. (2005). *Introduction to Neural Network with Java.* Chesterfield: Heaton Research Inc.

Hu, J., Xiong, C., Shu, J., Zhou, X. & Zhu, J. (2009). *A Novel Text Clustering Method Based on TGSOM and Fuzzy K-Means,* First International Workshop on Education Technology and Computer Science

IBM Corp.(2010). IBM® Rational® ClearCase® offers complete software configuration management. In *Rational ClearCase.* Retrieve Feb 26, 2011 from **http://www-01.ibm.com/software/awdtools/clearcase/**

Matteucci, M. (n.d.). K-Means Clustering. In *A Tutorial on Clustering Algorithm.* Retrieved Feb 25, 2011 from http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

Norsaadah Zakaria, Jamil Salleh, Mohd Nasir Taib, Tan, Y. Y. & Wah, Y.B., (2008). *Using Data Mining Technique to Explore Anthropometric Data towards the development of Sizing System.* Malaysia: Universiti Teknologi Mara

NeuroDimension Inc (2010). What is Neural Network?. In NeuroSolution. Retrieved Feb 25, 2011, from http://www.nd.com/neurosolutions/products/ns/whatisNN.html

Pan, D. (2009). *A Formal Framework for Data Mining Process Model,* Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications

Ping, O. A. (2009). *Analysis of Bankruptcy using data mining approach.* Unpublished master's thesis, Universiti Utara Malaysia, Kedah.

Fadzilah Siraj & Abdoulha, M.A.(2010). Mining Enrolment Data Using Predictive and Descriptive Approaches. In K. Funatsu & K. Hasegawa (Eds.), *Knowledge-oriented application in Data mining.*(pp. 53-72). Croatia: InTech

Flom, P. (n,d). Multinomial Logistic Regression. In Statistical Analysis Consulting. Retrieved Feb 25, 2011, from

http://www.statisticalanalysisconsulting.com/statistical-analysis-for-dissertations/research-methods/multinomial-logistic-regression/

Porter, Michael E. (1979). *How competitive forces shape strategy*. Harvard business review, 57(2): 137-145.

Rahm, E & Do,H. H. (Dec 2000). *Data Cleaning: Problems and Current Approaches,* Germany:University of Leipzig

Selbaş, R., Şencan, A. & Küçüksille, E.U.(2010). Data Mining Methods for Energy System Aplications. In K. Funatsu & K. Hasegawa (Eds.), *Knowledge-oriented application in Data mining*.(pp. 339-352). Croatia: InTech

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining, *Journal of Data Warehousing*, vol. 5, pp. 13–19

SPSS Inc. (2003). *SPSS Training: Introduction to Clementine.* *Retrieved 29 Jan 2011 from* http://homepage.univie.ac.at/marcus.hudec/Lehre/WS%202006/Methoden%20DA/IntroClem.pdf

SPSS Inc. (2007). *SPSS Neural Network 16.0.* *Retrieved 20 Feb 2011 from* http://www.uni-muenster.de/imperia/md/content/ziv/service/software/spss/handbuecher/englisch/spss_neural_network_16.0.pdf

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications*, New York: Springer-Verlog Berlin Heidelberg

Tan, P.N., Steinbach, M. & Kumar, V., (2006). *Introduction to Data Mining*, Addison-Wesley

UCLA: Academic Technology Services, Statistical Consulting Group (n.d.). *Annotated SPSS Output Multinomial Logistic Regression*, Retrieved 20 Feb 2011 from http://www.ats.ucla.edu/stat/spss/output/mlogit.htm

Webb, K. (2010). *Cluster Analysis*. Retrieved 23 Feb 2011, from http://www.cob.sjsu.edu/webb_k/B231A/ClusterR.doc

Wikipedia (2011). Self- Organizing map. Retrieved 25 Feb 2011, from http://en.wikipedia.org/wiki/Self-organizing_map

Yuan, W., & Yihua, Z., (2009). *Research on Classification and Subdivision Model of Telecom Rural Channel Based on Clustering Analysis.* International Conference on Information Management, Innovation Management and Industrial Engineering

Yun, C.H., Chuang, K.T., & Chen, M.S. (n.d). *An Efficient Clustering Algorithm for Market Basket Data Based on Small Large Ratios*, Taiwan : National Taiwan University

Zeng, H., &, Pan, D. (2010). *A Knowledge Discovery and Data Mining Process Model in E-Marketing.8th World Congress on Intelligent Control and Automation held on July 6-9 2010 at Jinan, China*

Zhou, J. (2004). *Comparing Regularized B-spline Neural Network, Multilayer Perceptron and Boosted-CART on Two Problems of Heart Arrhythmia Diagnosis*. Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco

Appendix A:

DATA MAPPING ON VARIABLES VALUE SUBSTITUTION

Operation_tier_one	Rep Value
(Blank)	0
Add	1
Application	2
End User Computing	3
Failure - Application	4
Hardware	5
Investigation	6
Lab Engineering	7
Midrange Computing	8
Modify	9
Other	10
Process	11
Remove	12
Repair	13
Support Services	14

Operation_tier_two	Rep Value
(Blank)	0
Account	1
Application	2
Business	3
CSC Internal Applications	4
Desktop	5
Desktop - Onsite	6
Desktop - Remote	7
Engineering	8
Hardware	9
Helpdesk	10
Operations	11
Other	12
Process	13
Software	14
Standard	15
Tools	16
Unix / Linux	17
Wintel	18

Operation_tier_three	Rep Value
(Blank)	0
Access	1
Add	2
Application	3
Business	4
Change	5
Configure	6
Create	7
DB	8
Delete	9
Failure	10
Install	11
Install/Reinstall	12
Modify	13
Other	14
Patch	15
Personal Computer	16
Rebuild	17
Reconfig	18
Remove	19
Repair	20
Raplace	21
Replication	22
Restart	23
Restore	24
Service	25
Status	26
Surplus	27
Upgrade	28

Product_tier_one	Rep Value
(Blank)	0
Account	1
Application	2
File Share	3
Hardware	4
Infrastructure	5
Networking	6
Other	7
Permissions	8
Software	9
Support Services	10

Product_tier_two	Rep Value
(Blank)	0
Acrobat	1
Apache	2
Application	3
BlackIce	4
Clear Orbit	5
ClearCase	6
ClearCase Multisite	7
Clearquest	8
Electric Cloud	9
ElectricCloud	10
Engineering	11
Framemaker	12
Klocwork	13
Microsoft	14
Midrange	15
Minitab	16
NA	17
NIS / DNS	18
Oracle	19
Other	20
PC Support	21
Perl	22
Quick Test Professional	23
Rational Rose	24
Rational Rose RT	25
Rose	26
Server	27
Services	28
Switch	29
Unix	30
Unix / Linux	31
VC++ for Windows	32
Web	33
Web Server	34
WEBTOOLS	35
Windows	36

Product_tier_three	Rep Value
(Blank)	0
-None-	1
Account	2
All	3
Asia	4
CASE	5
ECC Clearcase	6
Other	7
Project & Resource Management	8
Repair	9
Third Party	10
Uplink	11
Windows	12

Appendix B:

SCREENSHOT OF THE RAW DATA

	A	B	C	D	E	F	G	H	I	J	K	L
1	OPERATIONAL_TIER_ONE	OPERATION1	OPERATIONAL_TIER_THREE	OPERATION3	OPERATIONAL_TIER_TWO	OPERATION2	PRODUCT_TIER_ONE	PRODUCT1	PRODUCT_TIER_THREE	PRODUCT3	PRODUCT_TIER_TWO	PRODUCT2
122	Repair	13	Application	3	Software	14	Application	2	CASE	5	Engineering	11
123	Repair	13	Application	3	Software	14	Application	2	CASE	5	Engineering	11
124	Repair	13	Application	3	Software	14	Infrastructure	5		0	PC Support	21
125	Repair	13	Application	3	Software	14	Application	2	CASE	5	Engineering	11
126	Modify	9	Application	3	Software	14	Application	2	Project & Resource	8	Services	28
127	Repair	13	Application	3	Software	14	Application	2	CASE	5	Engineering	11
128	Repair	13	Application	3	Software	14	Application	2		0	Engineering	11
129	Investigation	6	Business	4	Process	13	Infrastructure	5		0	Midrange	19
130	Modify	9	Business	4	Process	13	Application	2	CASE	5	Engineering	11
131	Remove	12	Business	4	Process	13	Application	2		0	Engineering	11
132	Repair	13	Business	4	Process	13	Application	2	CASE	5	Engineering	11
133	Application	2	Change	5	Desktop	5	Application	2	- None -	1	Clear Quest	5
134	Application	2	Change	5	Engineering	8	Other	7		0	Other	20
135	End User Computi	3	Change	5	Wintel	16	Other	7		0	Other	20
136	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
137	Application	2	Configure	6	Engineering	8	Software	9	Third Party	11	Application	3
138	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase Midrange	7
139	Application	2	Configure	6	Engineering	8	Application	2	- None -	1	ClearCase	6
140	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
141	Application	2	Configure	6	Engineering	8	Application	2	- None -	1	Rational Rose RT	25
142	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
143	Application	2	Configure	6	Engineering	8	Application	2	- None -	1	ClearCase	6
144	Application	2	Configure	6	Desktop	5	Application	2	- None -	1	ClearCase	6
145	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
146	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
147	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
148	Application	2	Configure	6	Engineering	8	Application	2	- None -	1	ClearCase	6
149	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
150	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
151	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
152	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
153	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
154	Application	2	Configure	6	Engineering	8	Application	2	- None -	1	Rational Rose RT	25
155	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
156	Application	2	Configure	6	Engineering	8	Application	2		0	Rational Rose RT	25
157	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6
158	Application	2	Configure	6	Engineering	8	Application	2		0	ClearCase	6